**Step 1 : Ollama installation**

* Download Llama from <https://github.com/ollama/ollama>
* Install on the system
* Ollama run ollama3

A screenshot of a computer program

Description automatically generated

Now you can connect to it and ask your questions

**Step 2: Use Curl**

#### In this guidance <https://github.com/ollama/ollama/blob/main/docs/api.md#generate-a-completion>

#### you can find different models of Curl.

1. Go to Terminal
2. Write proper Curl. For example the Request (No streaming) is used:

curl http://localhost:11434/api/generate -d '{

"model": "llama3",

"prompt": "What is the capital city of Canada?",

"stream": false

}'

* **curl**: Initiates a HTTP request to the specified URL.
* **http://localhost:11434/api/generate**: Specifies the URL where the request is sent. In this case, it's http://localhost:11434/api/generate, which is typically a local server running on your machine (localhost) on port 11434.
* **-d**: Specifies that the following string is the data to be sent with the POST request.
* **'{ ... }'**: This JSON string contains the data payload for the request. It includes:
  + "model": "llama3": Specifies the model to use for generating the response. In this case, it's using the llama3 model.
  + "prompt": "What is the capital city of Canada?": Provides the input prompt for the model. The model will generate a response based on this prompt.
  + "stream": false: Indicates whether to stream the response or get it all at once (false means get it all at once).

3- The response is a Json, which include:

* "model": "llama3": Specifies the model used.
* "created\_at": "2024-06-17T17:36:58.761975Z": Timestamp of when the response was generated.
* "response": "The capital city of Canada is Ottawa.": The generated response based on the prompt.
* "done": true: Indicates that the response generation is complete.
* "done\_reason": "stop": Reason for completing the response.
* "context": Internal details related to the response generation process.
* "total\_duration", "load\_duration", "prompt\_eval\_count", "prompt\_eval\_duration": Metrics related to how long it took to generate the response and evaluate the prompt.

A screenshot of a computer

Description automatically generated

Gemma:2b

* Download and Install Gemma:2b by

ollama pull gemma:2b

* Run Gemma:2b

ollama run gemma:2b

* Ask simple question

Use Curl:

curl http://localhost:11434/api/generate -d '{

"model": "gemma:2b",

"prompt": "What is the capital city of Canada?",

"stream": false

}'

A screenshot of a computer

Description automatically generated